

Administración Informática de Colecciones de datos Heterogéneos.

Aplicación a Museos de Ciencias Naturales.

Hugo Ramón¹, Raúl Champredonde², Mariano Merino³, Patricia Pesado⁴, Armando De Giusti⁵

III-LIDI. Instituto de Investigación en Informática LIDI⁶

Facultad de Informática

Universidad Nacional de La Plata

Resumen

Se expone la experiencia realizada en el análisis y desarrollo de un sistema para el almacenamiento y recuperación de datos provenientes de las piezas albergadas en las colecciones del Museo de Ciencias Naturales de La Plata.

Esta experiencia está inscripta en un proyecto internacional, que involucra a diferentes países y en particular a la Red Nacional de Colecciones Biológicas.

El trabajo que se presenta describe dos aspectos principales: la definición del modelo de datos que soporte múltiples colecciones con características heterogéneas y un desarrollo flexible para realizar actualizaciones y consultas sobre los datos ingresados. En particular la gestión de consultas debe respetar perfiles de usuarios (investigadores, docentes, responsables de la colección particular, alumnos, público en general).

Un elemento a considerar en la resolución del caso planteado es no sólo la complejidad y variedad de las colecciones, sino el número de registros involucrados (del orden de 2 millones en el caso del Museo de Ciencias Naturales de La Plata). Esto obliga a tratar de recuperar total o parcialmente datos sistematizados en forma diferente para cada tipo de colección.

Por otra parte, la diversidad biológica de los elementos a almacenar, la localización de los elementos y la ubicación distribuida de operadores y usuarios de la información, han hecho útil el desarrollo de un sistema de software basado en WEB, con niveles de acceso. Actualmente este sistema está en prueba con un número importante de datos, correspondiente a las colecciones de Mastozoología (mamíferos) y Aracnología

Palabras Claves: Bases de Datos, Modelo de datos, Museos de Ciencias Naturales, Colecciones Biológicas, Sistemas de Información Geográfica.

¹ Lic. en Informática, Profesor Adjunto DE, Facultad de Informática, UNLP, hramon@lidi.info.unlp.edu.ar

² Lic. en Informática, Profesor Adjunto DE, Facultad de Informática, UNLP, rchampre@lidi.info.unlp.edu.ar

³ Dr. Mariano Merino, CICPA, División Zoología Vertebrados, Museo de La Plata, UNLP, mlmerino@fcnym.unlp.edu.ar

⁴ Lic. en Informática, CICBA, Profesor Titular, Facultad de Informática, UNLP, ppesado@lidi.info.unlp.edu.ar

⁵ Investigador Principal CONICET, Profesor Titular DE, Facultad de Informática, UNLP, degiusti@lidi.info.unlp.edu.ar

⁶ Calle 50 y 115 Primer Piso, (1900) La Plata, Argentina, Teléfono/fax +54(221)4227707

<http://www.lidi.info.unlp.edu.ar>

Introducción

La observación siempre está ligada al proceso de construcción de conocimiento. Esto es el punto de partida de la ciencia, en la que la observación es el primer paso del método científico [Wilson 1952].

La obtención de datos o información a partir de una observación de cualquier objeto de estudio, ubicado en un tiempo y espacio determina un *registro* que caracteriza al objeto en ese momento y lugar.

Los registros biológicos constituyen colecciones, un conjunto de datos complejos y vitales que tienen un alto potencial para conservación del conocimiento, definición de nuevas investigaciones, repetición de experimentos, predicciones y toma de decisiones para la conservación y uso sostenible de la biodiversidad.

Cada registro se describe con atributos biológicos que lo caracterizan, además de referencias geográficas y datos de la persona que recolecta el ejemplar. Existen estándares para documentar registros biológicos, como el Darwin Core 2 [HUM 2003].

Es fundamental la transmisión de conocimiento entre Instituciones e Investigadores y para facilitar el intercambio es necesario establecer mecanismos que unifiquen los conceptos. La información debe estar debidamente estructurada para realizar el intercambio digital en forma ágil y segura.

En un museo de Ciencias Naturales se coleccionan, preparan, preservan, estudian y exhiben los objetos de la Naturaleza que el público observa en sus salas. No se presentan organismos vivos, sino disecados, embalsamados o taxidermizados de modo adecuado y este hecho separa a los museos de los jardines botánicos, zoológicos y similares. De este modo, los museos de Ciencias Naturales se convierten en custodios y preservadores de colecciones y, además, simultáneamente se transforman en instituciones dedicadas a estudiar e investigar la naturaleza.

En particular en el Museo de Ciencias Naturales de La Plata, existen 35 colecciones biológicas de las cuales, algunas realizan los registros en forma manual y otras en diferentes formatos (archivos planos, word, excel, dbase, access) con aproximadamente 2.400.000 registros.

Este panorama no permite el intercambio, control de los ejemplares, ni publicación, con lo cual las autoridades del Museo, determinan la necesidad de la estructuración de la información.

La información de los ejemplares de las colecciones se actualiza continuamente, debido a las investigaciones que permanentemente se realizan.

El Museo de La Plata integra la *Red Nacional de Colecciones*, a través de la cual se accedió a un proyecto financiado por *Global Biodiversity Information Facility (GBIF)*, para informatizar colecciones.

Existen otros modelos propietarios [CON 2003] y recomendaciones [HUM 2003] para el diseño de bases de datos de colecciones biológicas, que se tuvieron en cuenta para realizar el esfuerzo de un desarrollo *in-house*, que agrega grados de libertad al proceso que implementa la aplicación.

Características de las Colecciones

El Museo de La Plata posee más de 2 millones y medio de objetos de colección, algunos en exhibición y la mayoría cuidados y conservados en los depósitos de las distintas Divisiones Científicas en las que

está organizado. Las colecciones biológicas, especialmente, son una muestra representativa de la biodiversidad de nuestro país. Encontrándose representados todos los grupos taxonómicos que conforman esta biodiversidad, desde los grandes mamíferos marinos, como la ballena azul, hasta las algas microscópicas.

Las colecciones más simples están caracterizadas por tener un ejemplar al cual se le asigna toda la información,

En las colecciones más complejas, por ejemplo ficología, la información es asignada a una muestra donde se encuentran diversos ejemplares de numerosas especies. En la colección de parasitología el ejemplar puede encontrarse alojado en otro ejemplar u hospedador.

Justamente el problema radica en que el modelo de datos tiene que poder soportar a todas estas características en las colecciones, donde se observa que hay una disparidad en la información asociada a cada ejemplar o muestra.

Requerimientos Generales de la Aplicación

La aplicación en su forma más general facilita los procesos de captura, administración, generación y divulgación de información sobre biodiversidad.

Mediante un código único, la información asociada a cualquier ejemplar de cualquier colección puede ser accedida y administrada en forma sistemática. La única restricción es que el usuario tenga algún tipo de derecho sobre la colección.

Sobre cada ejemplar recolectado se almacenan atributos que identifican:

- Qué es (sistemática),
- Dónde (georeferenciación y datos de la división política),
- Cuándo,
- Quién lo recolectó,
- Bajo que condiciones (vegetación, parámetros ambientales)

Se trató de mantener un modelo general (eliminando las especificidades de las colecciones), implementando un mecanismo de atributos dinámicos (meta datos) para las colecciones que necesitan atributos particulares, como por ejemplo, en Mastozoología los datos correspondientes a peso, largo del cuerpo y altura en la cruz.

Incluye módulos para generación, consulta, edición y análisis de información taxonómica, geográfica, y de usos potenciales de plantas, insectos, moluscos, arácnidos y animales.

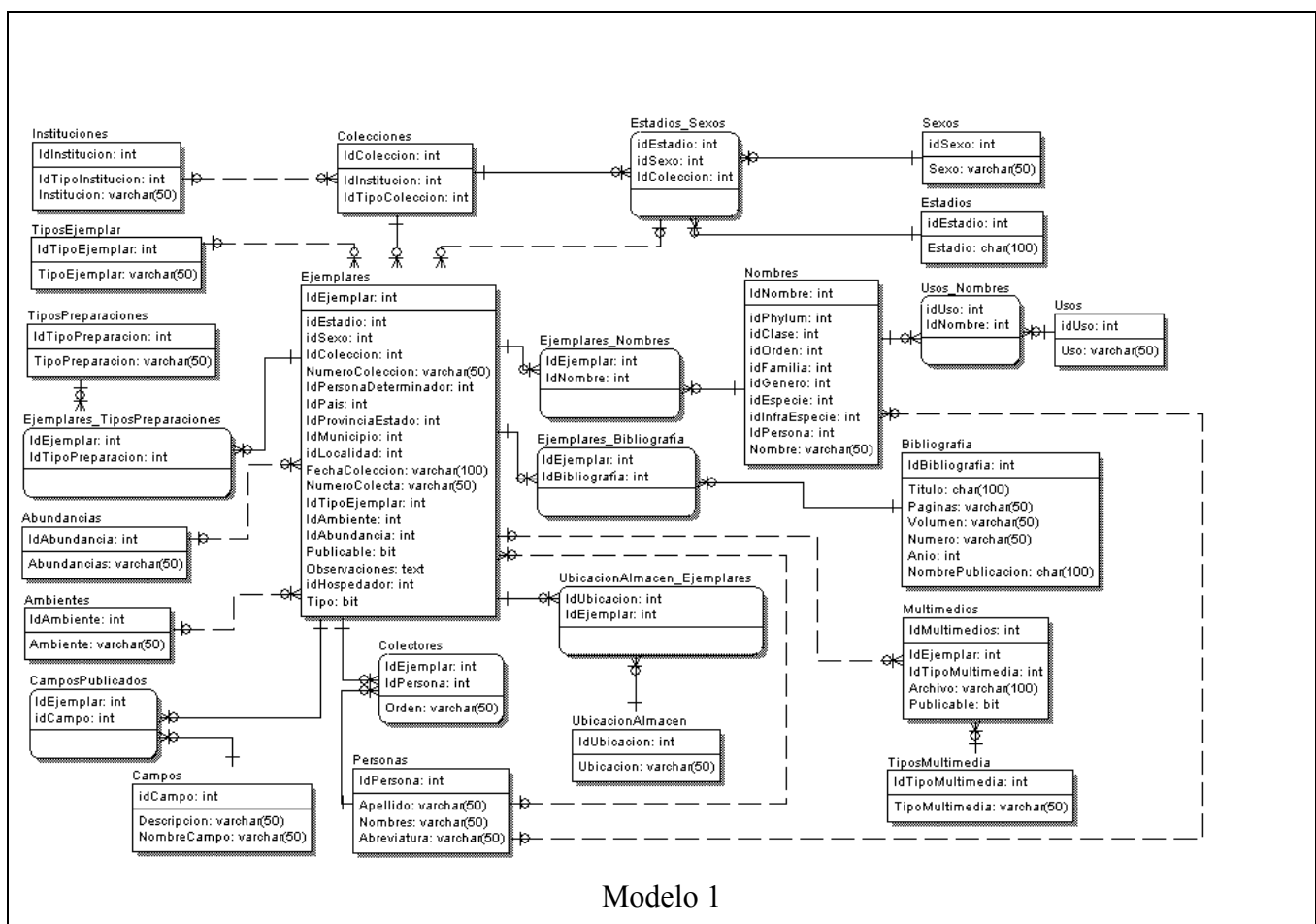
Es importante la logística de préstamos a otras Instituciones y el mantenimiento del preparado en donde se encuentran los ejemplares a través del uso de alarmas.

Mediante el uso de perfiles y usuarios la aplicación provee auditoría y trazabilidad de la información actualizada [WEB 1998].

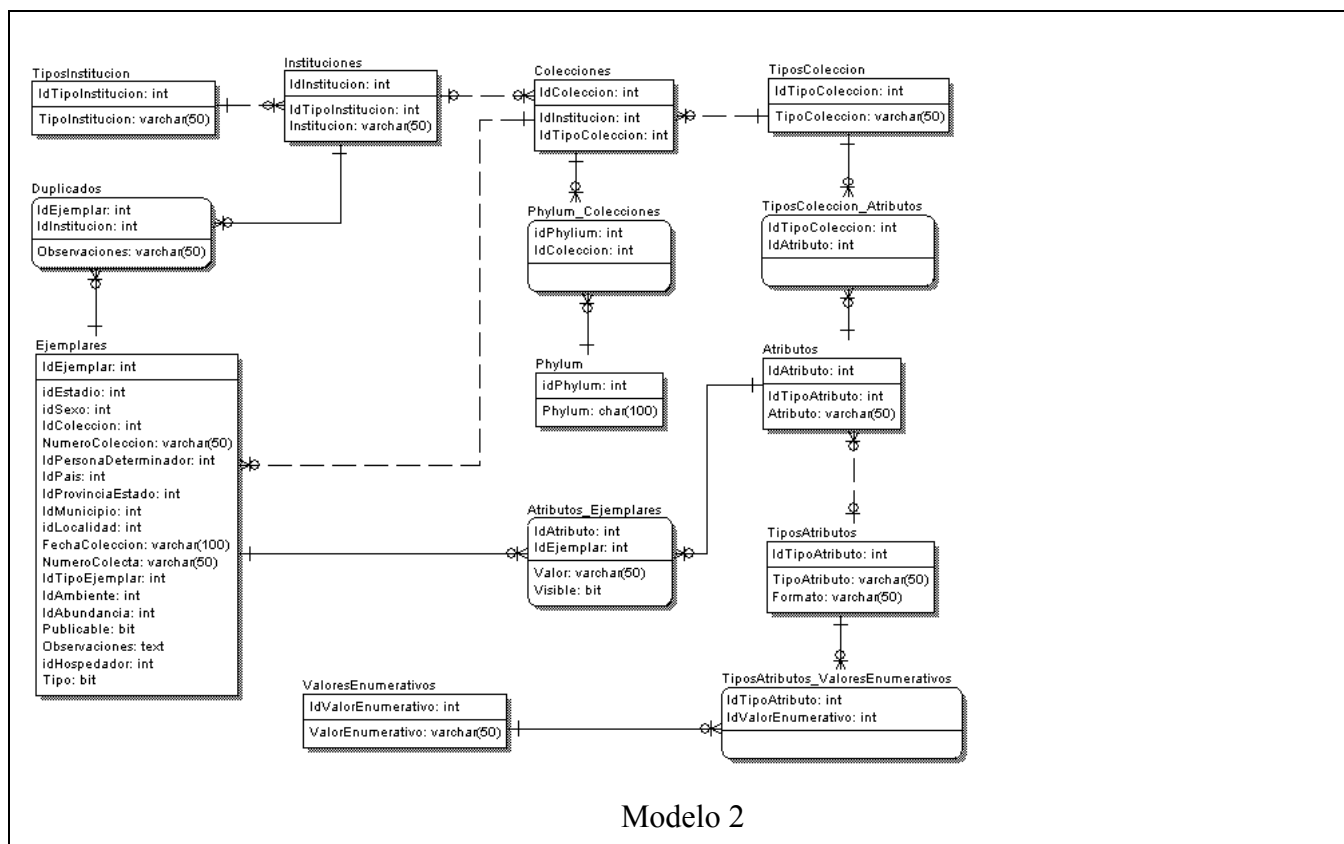
Características del Modelo de Datos

Para la definición del modelo de datos relacional [DATE 2001] se tuvo en cuenta el desarrollo definido en [CON 2003] y las recomendaciones de [HUM 2003] aplicando un modelo conceptual [HOFFER 2001] [PRESS 2002].

Se debió tener en cuenta que las características de los elementos dependen del tipo de colección, eso incluye la “sistemática” (identifica que es) y ciertos atributos comunes a todas las colecciones, por ejemplo información relacionada a bibliografía, colectores, preparados, etc. En la figura Modelo 1 se muestran los atributos correspondientes al submodelo de ejemplares, donde se puede apreciar como Estadios_Sexos depende del tipo de colección.



Los atributos particulares de cada colección, se representan mediante la definición de meta datos soportados por el modelo 2. El soporte permite para cada tipo de colección, definir un conjunto de atributos en forma dinámica, se pueden determinar el nombre del atributo (tabla atributo) y de que tipo es (incluso su formato de presentación) y hasta podrían definirse enumerativos. Esta forma de representar la información particular permite incluso incorporar ejemplares de las colecciones paleontológicas.



Conversión de Datos Existentes

La conversión de datos existentes es una tarea crítica y masiva debido a la diversidad de soporte de los registros actuales. Esto incluye archivos planos, planillas excel, archivos dbase, documentos word con información no normalizada.

Luego de definido el modelo se realizó en forma artesanal, colección por colección, el mapeo de los campos existentes a los del nuevo modelo, incorporando al curador de la colección en el proceso de conversión.

Hasta este momento se han convertido 3 colecciones y se está avanzando en la conversión de 4 adicionales.

Desarrollo actual y Resultados Obtenidos

Los requerimientos funcionales mínimos están implementados en una aplicación que se desarrolló aplicando tecnología de eXtreme Programming [BEC 2000] [PAU 2001] por un equipo de cuatro técnicos y dos usuarios (de las colecciones de Mastozoología e Invertebrados).

Se desarrolló una base de datos que al momento almacena 10000 elementos de las colecciones de Mastozoología y Aracnología. Las consultas definidas permiten la recuperación de los elementos almacenados en base a un acuerdo sobre los criterios definidos en cuanto a atributos y privacidad.

Actualmente se avanza en la conversión e ingreso de las colecciones Arácnida, Myriápoda, Moluscos, Ficología hasta llegar a almacenar los 20000 ejemplares requeridos por el proyecto GBIF.

Conclusiones

Se ha definido un modelo de datos y los requerimientos que soportan la dinámica en la administración de las Colecciones del Museo de La Plata, en forma general, sin distorsionar el modelo con particularidades. Estos requerimientos están implementados en un prototipo operacional.

Esto permite tener definida una base de datos completa y concreta para la digitalización de las colecciones de otros museos, en una base nacional orientada a la investigación sobre la biodiversidad y sus aplicaciones sobre el intercambio de información.

Trabajos Futuros

Está prevista como próxima fase la incorporación de ejemplares del Herbario en donde no existe registro electrónico y se analiza la conversión de datos de la colección de Ficología.

Esta base de datos relacional que almacenará los 2.400.000 ejemplares de las colecciones del Museo de La Plata con información de localización, permitirá integrarla a un sistema de información geográfica en donde los investigadores y las autoridades gubernamentales podrán obtener información detallada de la biodiversidad de una región de Argentina y con esa información definir experimentos y políticas.

Se debe integrar la información de las colecciones digitalizadas a través de GBIF a una base de datos distribuida [DIG 2001] como proveedor de datos o recursos.

Bibliografía

- [BEC 2000] Extreme Programming Explained: Embrace Change. Kent Beck. Addison-Wesley Longman, 2000.
- [CON 2003] COMisión NAcional para el conocimiento y uso de la Biodiversidad (CONABIO), Biótica 4.1 www.conabio.gob.mx.
- [DATE 2001] C. J. Date. Introducción a los Sistemas de Bases de Datos, Prentice may, 2001
- [DIG 2001] Distributed Generic Information Retrieval (DiGIR), <http://www.digir.net/>
- [GBI 1999] Global Biodiversity Information Facility (GBIF), <http://www.gbif.net>
- [HOFFER 2001] Modern Database Managment, J. Hoffer, M. Prescott, Prentice Hall, 2001.
- [HUM 2003] Instituto “Alexander von Humboldt”, Estándar para la documentación de registros biológicos, <http://www.humboldt.org.co>.
- [PAU 2001] Extreme Programing from a CMM Perspective. M. Paulk, IEEE Software. Vol 18 - 6, pp 19-26. 2001.
- [PRESS 2002] Ingeniería de Software. Un Enfoque Clásico. R. Pressman, Mc. Graw Hill, 2002.
- [WEB 1998] Information Systems Control and Audit, R. Weber, Prentice Hall, 1998.
- [WIL 1952] Wilson, E. Bright. An Introduction to Scientific Research, McGraw-Hill, 1952.